A

Major Project

On

PREDICTIVE ANALYSIS FOR SUPERMARKET SALES USING MACHINE LEARNING

(Submitted in partial fulfillment of the requirements for the award of Degree)

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING
BY

Aditya Singh(187R1A05C4)

Abhinav Dharipalli(187R1A05E0)

RamaKrishna Bollepally(187R1A05C6)

Under the Guidance of

D.MOUNIKA

(Assistant Professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR TECHNICAL CAMPUS

UGC AUTONOMOUS

(Accredited by NAAC, NBA, Permanently Affiliated to JNTUH, Approved by AICTE, New Delhi)

Recognized Under Section 2(f) & 12(B) of the UGCAct.1956,

Kandlakoya (V), Medchal Road, Hyderabad-501401.

2018-22

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled"**PREDICTIVE ANALYSIS FOR SUPERMARKET SALES USING MACHINE LEARNING**" being submitted by **ADITYA SINGH (187R1A05C4), ABHINAV DHARIPALLI(187R1A05E0), RAMAKRISHNA BOLLEPALLY(187R1A05C6)** in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by him/her under our guidance and supervision during the year 2021-22.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Mrs. D. MOUNIKA
Assistant Professor
INTERNAL GUIDE

Dr. A. Raji Reddy
DIRECTOR

Dr. K. Srujan Raju
HOD

EXTERNAL EXAMINER

Submitted for viva voice Examination heldon _____

ACKNOWLEGDEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We take this opportunity to express my profound gratitude and deep regard to myguide

Mrs.D.Mounika, Assistant Professor for his exemplary guidance, monitoring and constant encouragement throughout the project work.The blessing, help and guidance given by him shall carry us a long way in the journey of life on which we are about to embark.We also take this opportunity to express a deep sense of gratitude to Project Review Committee (PRC) Dr. **M. Varaprasad Rao, Mr. J. Narasimha Rao, Dr. T. S. Mastan Rao, Dr. Suwarna Gothane, Mr. A. Uday Kiran, Mr. A. Kiran Kumar, Mrs. G. Latha** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. K. Srujan Raju,** Head, Department of Computer Science and Engineering for providing encouragement and support for completing this project successfully.

Weare obliged to **Dr. A. Raji Reddy,** Director for being cooperative throughout the course of this project. We also express our sincere gratitude to Sri. **Ch. Gopal Reddy,** Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of thisproject.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

ADITYA SINGH(187R1A05C4)

ABHINAV DHARIPALLI(187R1A05E0)

RAMAKRISHNA BOLLEPALLY(187R1A05C6)

ABSTRACT

Future sales prediction is a crucial component of any organization. Accurate prediction of future sales assists the company in enhancing and managing business strategies as well as gaining a complete knowledge of the economy. Standard sales forecasts help companies in analyzing previous scenarios and applying customer purchase inferences to discover shortfalls and weaknesses prior to budgeting and planning for the coming year. A deep understanding of previous opportunities allows one to plan for future market demands and boost one's chances of success. Big Marts currently keep track of sales data for each individual item in order to predict future consumer demand and update inventory management. Engineers assist marts by estimating sales per product with the use of data science. Good forecasting allows products to be sold efficiently and stores to profit from them. Mining the data warehouse's data store is a common way to discover anomalies and general trends. The resulting data can be used by retailers such as Big Mart to forecast future sales volume using various machine learning techniques. For forecasting a company's sales, Linear regression, Polynomial regression, Xgboost and Random Forest techniques can be used to create a predictive model.

LIST OF FIGURES/TABLES

LIST OF SCREENSHOTS

TABLE OF CONTENT

1.INTRODUCTION

1. INTRODUCTION

1.1 PROJECT SCOPE

This project is titled "PREDICTIVE ANALYSIS FOR SUPERMARKET SALES BY USING MACHINE LEARNING". Currently, Big Marts keep track of each individual item's sales data in order to anticipate potential consumer demand and update inventory management.With the help of data science, engineers help the marts by predicting the sales per product. By good prediction the products can be sold efficiently and stores can generate good profits from them.Anomalies and general trends are often discovered by mining the data warehouse's data store. For retailers like Big Mart, the resulting data can be used to forecast future sales volume using various machine learning techniques. A predictive model can be developed using Xgboost, Linear regression, Polynomial regression, and Random Forest techniques for forecasting the sales of a business.

1.2 PROJECT PURPOSE

Much of the current work is focused in two major directions.Understanding the factors that effect the sales of a product and help in organizations in anticipating future demand, and Predicting the future demand of those products.

1.3 PROJECT FEATURES

Analyzing the past sales of the products. Understanding the factors that effect the sales of a product.Deriving inferences related to those sales.Predicting the future sales from the inferences derived.Help the businesses stock up/stock down products accordingly.

2.SYSTEM ANALYSIS

2. SYSTEM ANALYSIS

2. SYSTEM ANALYSIS

It is a process of collecting and interpreting facts, identifying the problems, and decomposition of a system into its components. System analysis is conducted for the purpose of studying a system or its parts in order to identify its objectives. It is a problem solving technique that improves the system and ensures that all the components of the system work efficiently to accomplish their purpose.Analysis specifies what the system should do.

2.1 PROBLEM DEFINITION

Day by day the competition in the market is increasing rapidly with the new players entering the market space with unique strategies to attract the customers. In such a competitive environment one shouldn't fall behind by not understanding the needs of the consumer. Predicting the future demand of any product and stocking them accordingly is an essential in every business organization. With a precise prediction one can achieve better customer retention and satisfaction and avoid over-stock and under-stock situations. Accurate forecast of the future sales help the organization develop a business plan or strategy according to the demand and current status of the market.

2.2 EXISTING SYSTEM

In the current big-mart, many products will be available but the owner or a salesmen cannot know which product will have high demand. The basic and foremost technique used in the sales is the statistical methods, which is also known as traditional methods, but these methods much more time for predicting the sales and also these meth -ods could not handle the non-linear data so to overcomethese problems in the traditional methods machine learning models are deployed.

2.2.1 LIMITATIONS OF EXISTING SYSTEM

- Waste of time and traveling cost.
- Shortage of products or over stock of products may occur.
- Does not work efficiently for many big cities.
- There is a need to maintain a large number of employees for surveying.

2.3 PROPOSED SYSTEM

In this project we are going to develop a machine learning model to predict the future demand of products from a particular mart. We are going to perform all the operations on big-mart dataset. We will try to perform data cleaning and apply classifiers like Linear Regression, Random Forest, Decision Trees, XGboost and other machine learning algorithms.

We will check the accuracy of our developed machine learning model using python libraries. The model will predict the both minimum and maximum demand for the product. This approach may be useful and overcome the disadvantages of the existing approach.

2.3.1 ADVANTAGES OF PROPOSED SYSTEM

- No Waste of time and No traveling cost
- No Shortage of products or over stock of products may occur.
- Work efficiently for many big cities.
- There is a no need to maintain a large number of employees for surveying.

2.4 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system

analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. Three key considerations involved in the feasibility analysis.

- Economic Feasibility
- Technical Feasibility
- Social Feasibility

2.4.1 ECONOMIC FEASIBILITY

Development of this application is highly economically feasible. The organization needed not spend much money for the development of the system already available. The only thing is to be done is making an environment for the development with an effective supervision. If we are doing so, we can attain the maximum usability of the corresponding resources. Even after the development, the organization will not be in condition to invest more in the organization. Therefore, the system is economically feasible.

2.4.2 TECHNICAL FEASIBILITY

We can strongly say that it is technically feasible, since there will not be much difficulty in getting required resources for the development and maintaining the system as well. All the resources needed for the development of the software as well as the maintenance of the same is available in the organization here we are utilizing the resources which are available already.

2.4.3 BEHAVIOURAL FEASIBILITY

Whatever we think need not be feasible. It is wise to think about the feasibility of any problem we undertake. Feasibility is the study of impact, which happens in the organization by the development of a system. The impact can be either positive or negative. When the positives nominate the negatives, then the system is considered feasible. Here the feasibility study can be performed in two ways such as technical feasibility and Economical Feasibility.

2.5 HARDWARE & SOFTWARE REQUIREMENTS

2.5.1 HARDWARE REQUIREMENTS

Hardware interfaces specifies the logical characteristics of each interface between the software product and the hardware components of the system. The following are some hardware requirements.

- System: Intel i3
- Hard disk: 16GB and above
- Input devices: Keyboard, mouse
- Ram: 4GB and above

2.5.2 SOFTWARE REQUIREMENTS

Software Requirements specifies the logical characteristics of each interface and software components of the system. The following are some software requirements,

- Operating system: Windows 7,8,10,11
- Coding language: Python
- Tool: Google colab/jupyter notebook/pycharm

3.ARCHITECTURE

3. ARCHITECTURE

3.1 PROJECT ARCHITECTURE

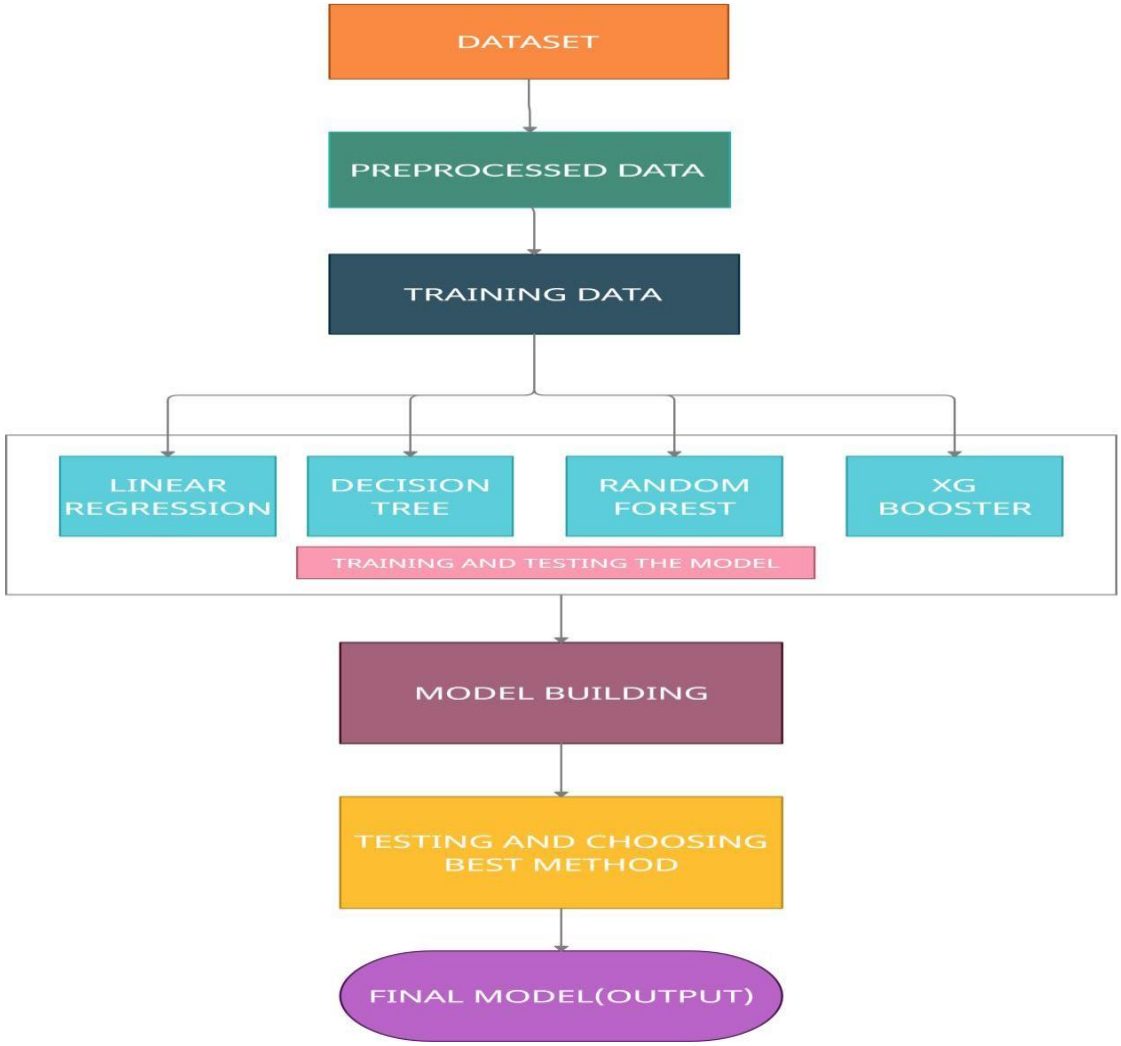


Figure 3.1: Project architecture of PREDICTIVE ANALYSIS FOR SUPERMARKET SALES BY USING MACHINE LEARNING

3.2 MODULES DESCRIPTION

The project is divided into 3 modules. The modular approach of the project is shown below in sequential manner

Module 1: Data Preprocessing

This module mainly deals with data preprocessing. The dataset may contains null values or outliers. The noise in the data may effects the accuracy of the model.

Attributes in dataset:-

Item identifier

Item weight

Item type

Item mrp

Outlet Identifier

Outlet size

Outlet establishment year

Outlet location type

Item outlet sales

Outlet type

Module 2: Splitting and Training

- Whole dataset split in to 2 parts Testing dataset(test.csv) and training dataset(train.csv).
- We will build various machine learning models with many ML algorithms like linear regression, Random forest, Xgboost. Both testing and training will be

performed using all of these algorithm.

Module 3: Final Output

- In this module we mainly deals with comparison of results obtained by above machine learning algorithms.
- Finally the algorithm with maximum accuracy and less error rate will be chosen and this module also deals with data visualization all the required graphs will be displayed to the user.

3.3 USE CASE DIAGRAM

A use case is a set of scenarios that describing an interaction between a user and a system. A use case diagram displays the relationship among actors and use cases. The two main components of a use case diagram are use cases and actors.

An actor is represents a user or another system that will interact with the system you are modeling. A use case is an external view of the system that represents some action the user might perform in order to complete a task.

Contents:

- Use cases
- Actors
- Dependency, Generalization, and association relationships
- System boundary

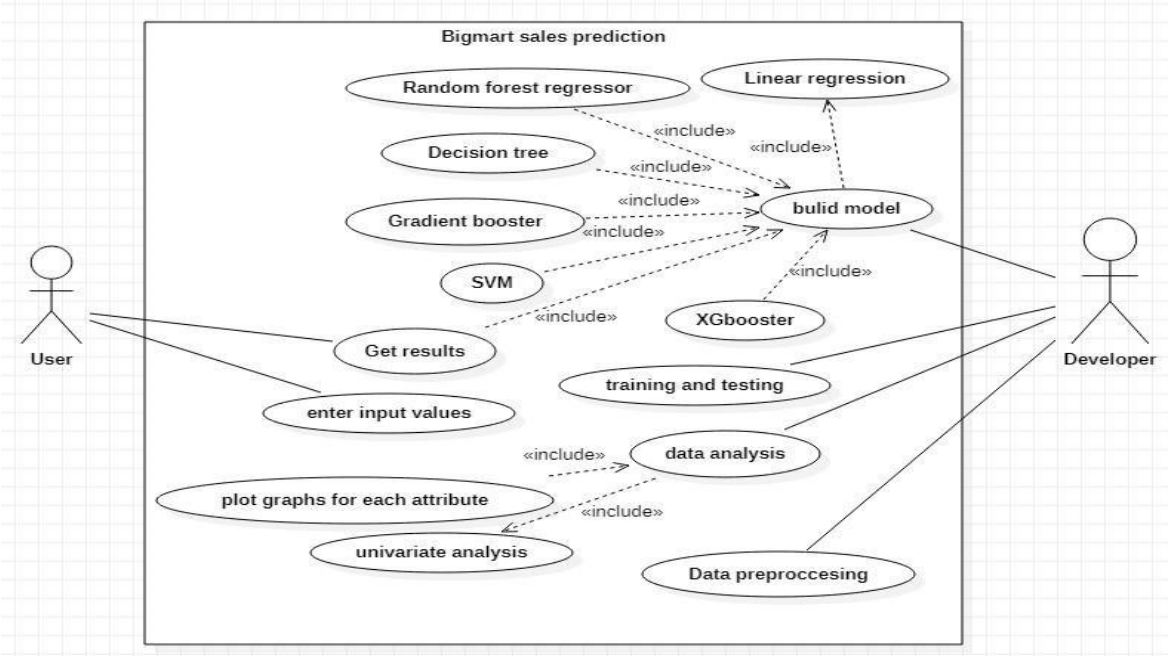


Figure3.2 Usecase diagram for PREDICTIVE ANALYSIS FOR SUPERMARKET SALES BY USING MACHINE LEARNING

3.4 CLASS DIAGRAM

Class diagrams are widely used to describe the types of objects in a system and their relationships. Class diagrams model class structure and contents using design elements such as classes, packages and objects. Class diagrams describe three different perspectives when designing a system, conceptual, specification, and implementation. These perspectives become evident as the diagram is created and help solidify the design. Class diagrams are arguably the most used UML diagram type. It is the main building block of any object oriented solution. It shows the classes in a system, attributes and operations of each class and the relationship between each class.

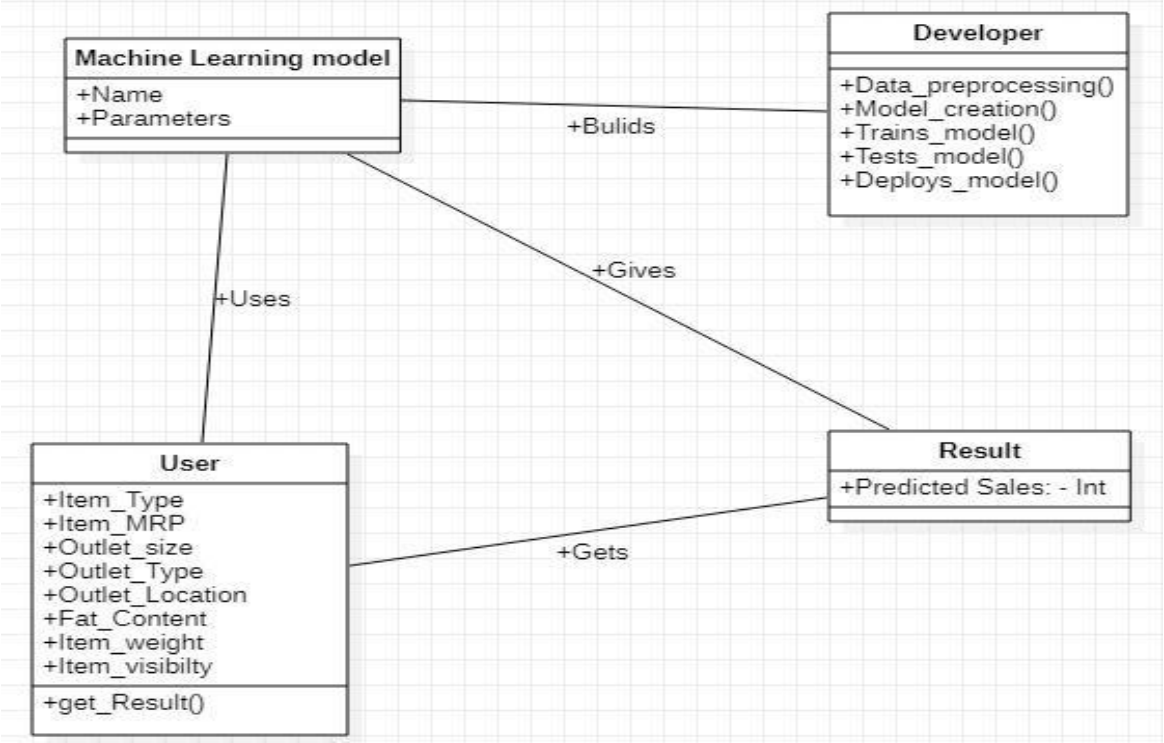


Figure3.3 Class diagram for PREDICTIVE ANALYSIS FOR SUPERMARKET SALES BY USING MACHINE LEARNING

3.5 SEQUENCE DIAGRAM

Sequence diagrams in UML shows how object interact with each other and the order those interactions occur. It's important to note that they show the interactions for a particular scenario. The processes are represented vertically and interactions are shown as arrows.

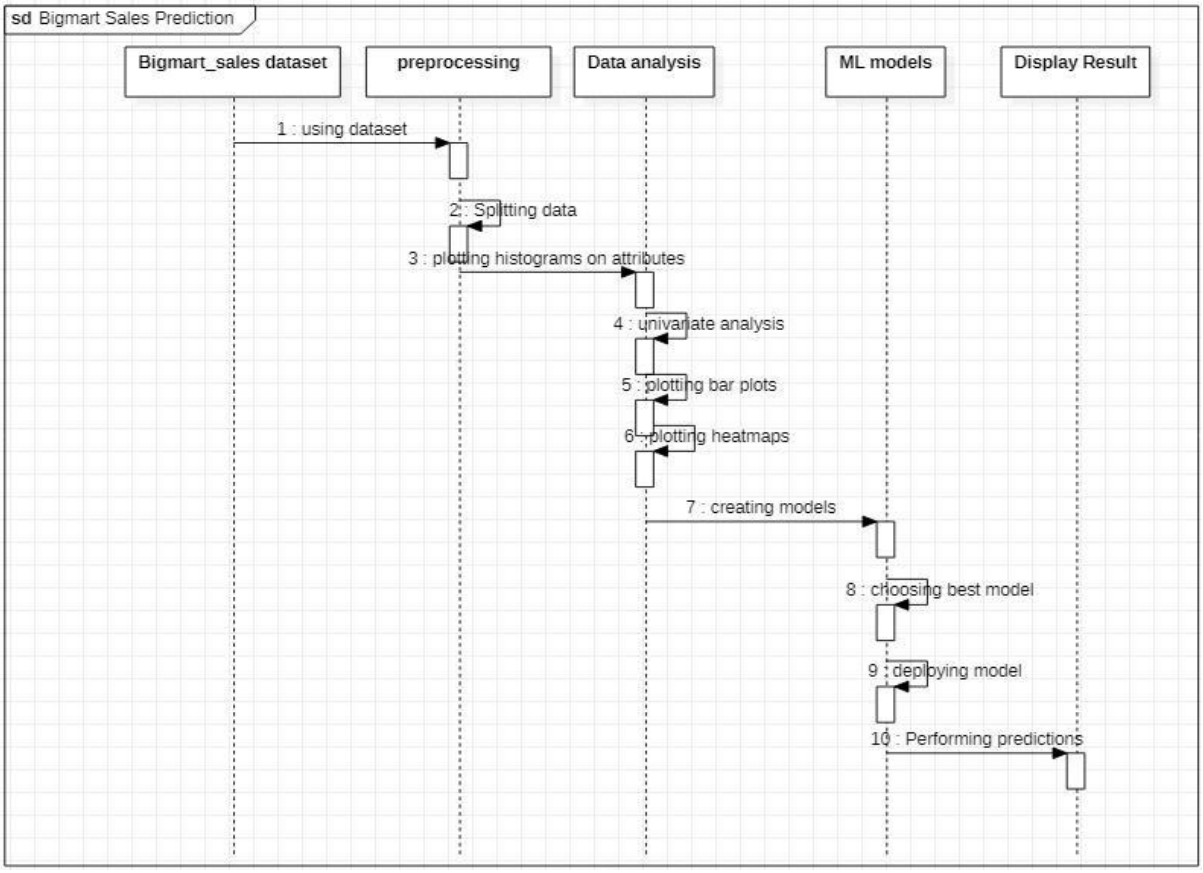


Figure3.4 Sequence diagram for PREDICTIVE ANALYSIS FOR SUPERMARKET SALES BY USING MACHINE LEARNING

3.6 ACTIVITY DIAGRAM

It describes about flow of activity states.

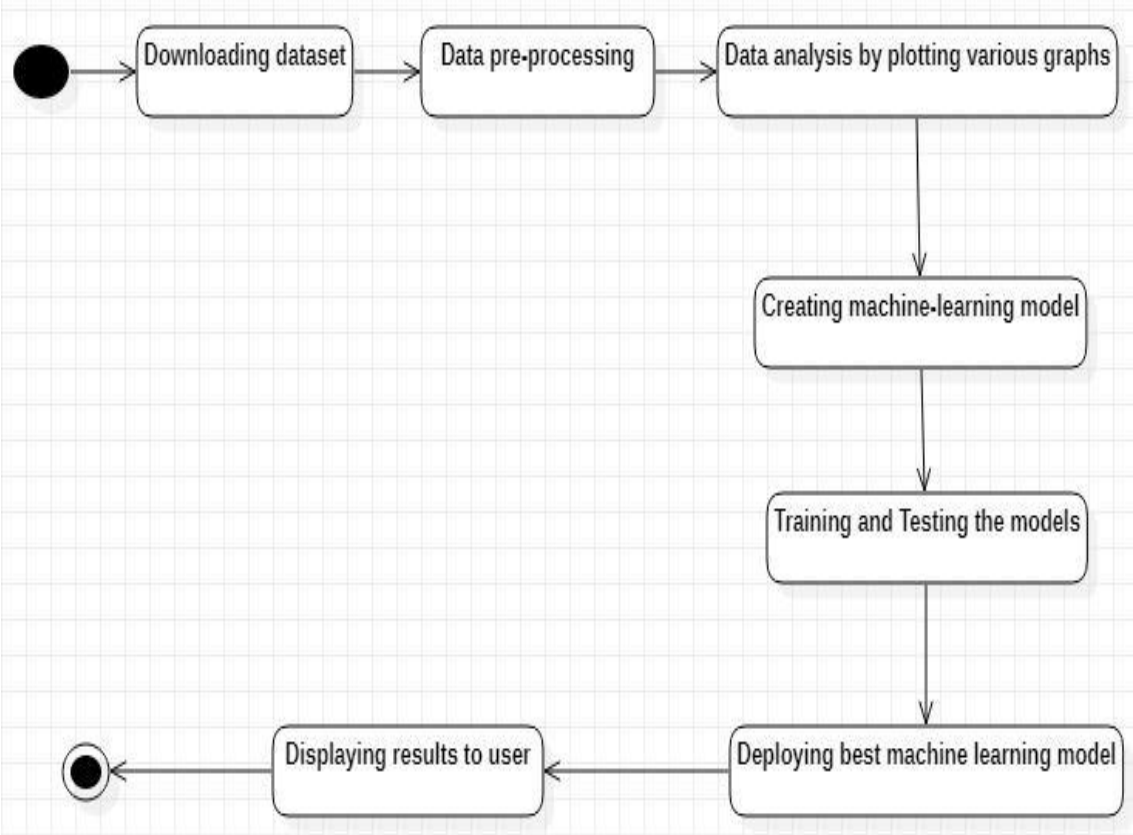


Figure3.5: Activity diagram for PREDICTIVE ANALYSIS FOR SUPERMARKET SALES BY USING MACHINE LEARNING

4.IMPLEMENTATION

4. IMPLEMENTATION

4.1 SAMPLE CODE

```
  import·numpy·as·np
 import·pandas·as·pd
 import·seaborn·as·sns
 import·matplotlib.pyplot·as·plt
 %matplotlib·inline
 import warnings
warnings.filterwarnings('ignore')
# Measuring Accuracy
from sklearn.metrics import accuracy_score, r2_score, mean_squared_error
from sklearn.model_selection import cross_val_score
from sklearn import metrics
df=pd.read_csv("/content/Bigmart1.csv")
df.head()
df.describe(include='all')
def imputeItemWeight(cols):
identifier = cols[0]
weight = cols[1]
if pd.isnull(weight):
return avgWeight[identifier]
else:
return weight
df['Item_Weight'] = df[['Item_Identifier', 'Item_Weight']].apply(imputeItemWeight, axis=1)
df.isnull().sum()
```

Train Test Split

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=101)
```

Building regression model

```
from sklearn.linear_model import LinearRegression

lm = LinearRegression(normalize=True)

lm.fit(X_train, y_train)

LinearRegression(normalize=True)
```

Prediction and evaluation

```
predict = lm.predict(X_test)

print('Score: ')

lm.score(X_test, y_test)

#r2_score(y_train, lm.predict(y_test))

from sklearn import metrics

print('MAE:', metrics.mean_absolute_error(y_test, predict))

print('MSE:', metrics.mean_squared_error(y_test, predict))

print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predict)))
```

Random forest regressor

```
from sklearn.ensemble import RandomForestRegressor

rf = RandomForestRegressor()

rf.fit(X_train, y_train)

RandomForestRegressor()
```

Prediction and evalution

```
rfpredict = rf.predict(X_test)

print('Score: ')

rf.score(X_test, y_test)

print('MAE:', metrics.mean_absolute_error(y_test, rfpredict))

print('MSE:', metrics.mean_squared_error(y_test, rfpredict))

print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, rfpredict)))
```

Decision tree model

```
# Fitting Decision Tree Regression to the dataset

from sklearn.tree import DecisionTreeRegressor

regressor = DecisionTreeRegressor(max_depth=15,min_samples_leaf=300)

regressor.fit(X_train, y_train)

DecisionTreeRegressor(max_depth=15,     min_samples_leaf=300)

# Predicting the test set results

y_pred = regressor.predict(X_test)

y_pred
array([1471.86266787, 2468.9944625 , 1808.6857972 , ..., 1471.86266787,
3139.96117009, 4071.19806567])

tree_accuracy = round(regressor.score(X_train,y_train),2)

tree_accuracy
```

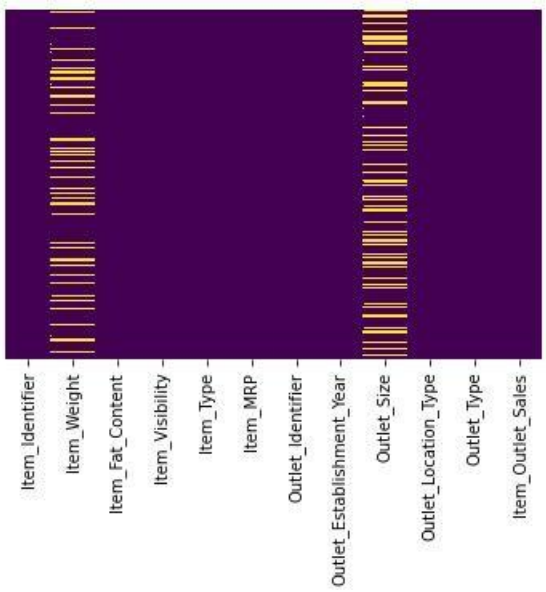
```
r2_score(y_train, regressor.predict(X_train))

print("RMSE : %.4g" % np.sqrt(metrics.mean_squared_error(y_train, regressor.predict(X_train))
```

XGBooster

```
from xgboost import XGBRegressor

regressor=XGBRegressor()

regressor.fit(X_train,y_train)

WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now d

XGBRegressor()

# predicting the test set results

y_pred = regressor.predict(X_test)

print(y_pred)

# Calculating the RMSE Score

mse = mean_squared_error(y_test, y_pred)

print("RMSE :", np.sqrt(mse))

[1377.7054 2561.2925 1380.0808 ... 1434.3055 3198.0376 3814.3193]

RMSE : 1024.0522593480985

regressor.score(X_test, y_test)
```

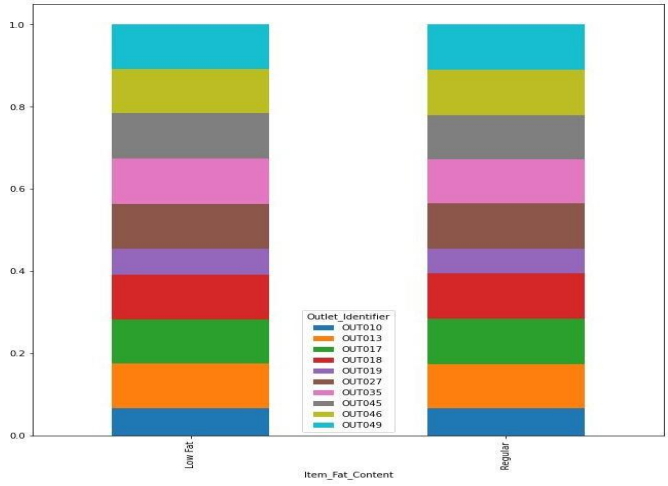
5.SCREENSHOTS

5.1 HEATMAP ON ATTRIBUTES



Screenshot 5.1: Heatmap on attributes

5.2 outlet_identifier vs fat_content

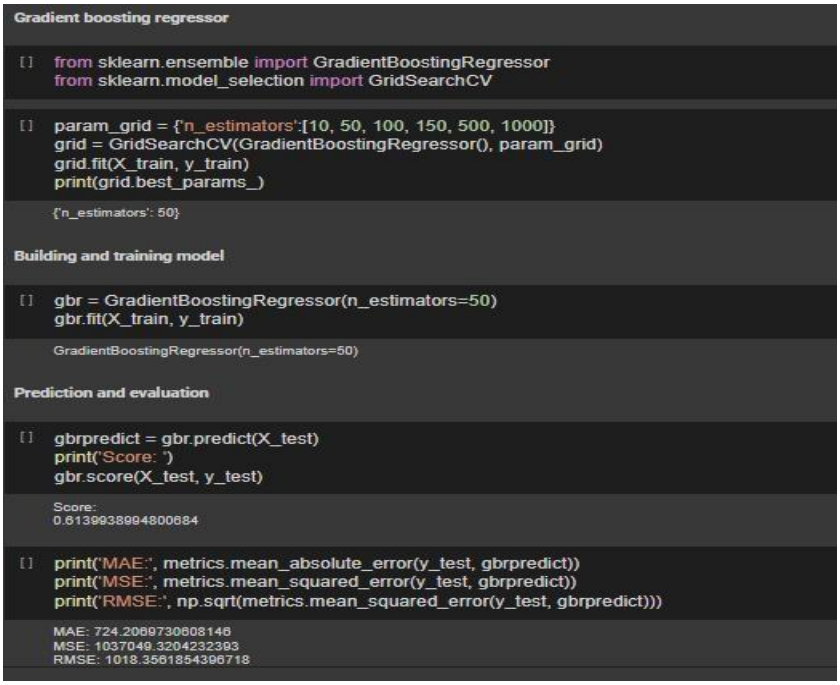


Screenshot 5.2:outlet_identifier vs fat_content

5.3 XGBOOSTER

Screenshot 5.3: XGBooster

5.4 GRADIENT BOOSTER



Screenshot 5.4: Gradient booster

6.TESTING

6. TESTING

6.1 INTRODUCTION TO TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

6.2 TYPES OF TESTING

6.2.1 UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

6.2.2 INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the

components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

6.2.3 FUNCTIONAL TESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes**.**

6.3 TEST CASES

6.3.1 CLASSIFICATION

Test case id	Test case name	Purpose	output
1	Linear regressionTest-1	To check the performance of the model	Less accuracy
2	Decision Tree Test-2	To check the performance of the model	Moderate accuracy
3	Random Forest Test-3	To check the performance of the model	Average accuracy
4	XGBoost Regressor Test-4	To check the performance of the model	More Accuracy

7.CONCLUSION

7. CONCLUSION & FUTURE SCOPE

7.1 PROJECT CONCLUSION

In present era of digitally connected world every shopping mall desires to know the customer demands beforehand to avoid the shortfall of sale items in all seasons.The Project focuses on predicting the future demand of the products in a mart using machine learning techniques.he store managers can make use of this system to know the future demand and stock up goods in their market accordingly.This system avoids situations of over stock and under stock if predicted rightly.currently working on deploying this model onto a website for better practical usage.

7.2 PROJECT FUTURE SCOPE

Till now we have completed all the data pre-processing and data visualizationpart. We need to explore on more machine learning models in future to attain good accuracy and precision. We also are trying to create good front end to our project.Such that user can easily interact with our application.We are planning to write and publish the paper on our project. We are stilling in planning phase we will start writing the paper with in short time. We are planning to add more graphs and plots in data visualization so that we can study every attribute and can decide how to deal with machine learning models.

8.BIBLIOGRAPHY

8. BIBLIOGRAPHY

8.1 REFERENCES

[1] https://www.kaggle.com/yonatanrabinovich/bigmart-sales-data-analysis-and-prediction sales data for 1559 products across 10 stores in different cities

[2] Beheshti-Kashi, S., Karimi, H.R., Thoben, K.D., L¨utjen, M., Teucke, M.: A survey on retail sales forecasting and prediction in fashion markets. Systems Science & Control Engineering 3(1), 154–161 (2015)

[3] Blog: Dataaspirant, “HOW THE RANDOM FOREST ALGORITHM WORKSINMACHINELEARNING”,(URL:https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/.).

[4] Mohit Gurnani, Yogesh Korke, Prachi Shah, Sandeep Udmale, Vijay Sambhe, Sunil Bhirud, “Forecasting of sales by using fusion of machine learning techniques”,2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), IEEE, October 2017.

[5] Kadam, H., Shevade, R., Ketkar, P. and Rajguru.: “A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression.” (2018).

8.2 WEBSITES

[1] www.medium.com

[2] www.edureka.co/blog/python-libraries/

[3] www.kaggle.com